EE102 Project

For the final project, you could work with at most two partners (i.e., ***up to*** three students in one group). The purpose of the project is for you to gain experience in applying the statistics knowledge/techniques to a real data set of interest to you. The project contains two parts:

A. Problem Definition and Data Collection

B. Modeling and Analysis

For modeling and analysis, you can use any of the following programming languages: Matlab, Python, C, Excel, etc (you are free to choose any of those languages).

Step1. Decide a question of interest

Examples of questions of interest are as follows:

What factors are related to a student's GPA?

What properties of a college are related to its rank in the U.S. News and World Report rankings in a particular year?

What properties of a country are related to the life expectancy of that country in a particular year?

What properties of a house are related to its price in a particular year?

(Your project topics are not limited to the above examples.)

Decide your project topic and decide the output variable of your project (in the above examples, the output variable is a college's US news ranking in year 2016 (2016 is a particular year), country's life expectancy in year 2012 (2012 is a particular year), house price in year 2017 (2017 is a particular year), respectively).

Then decide the input variables of your project. For example, if the question you select is the country's life expectancy, then the possible properties of a country (that may affect life expectancy) may include its GDP, the HIV rate, unemployment rate, etc. If you select the house price project, the properties of a house (that may affect its price) may include the number of bedrooms, the house built year, the number of bathrooms in the house, etc. These properties or factors that may affect your output variable are the input variables of your project.

Requirements:

1. **Consider 5 input variables**
2. **The input/output variables' values should be numbers (real numbers or integer numbers)**
3. **It is better to have independent input variables (i.e., the five input variables are mutually independent)**

Step 2. Collect the Real Dataset

The dataset should contain ***at least*** 100 observations. For example, if your project is house price in year 2012 at San Jose, you need to collect at least 100 houses' prices at San Jose in year 2012, and collect these houses' properties (input variables), such as number of bedrooms, number of bathrooms, etc. If your project is country's life expectancy in year 2010, you need to collect at least 100 countries' life expectancy in year 2010, and collect these countries' properties, such as GDP, HIV rate, unemployment rate, etc.

Create a form (e.g., Excel) including these data set.

You can find some data set available online at https://www.gapminder.org/data/

http://archive.ics.uci.edu/ml/datasets.html?format=&task=reg&att=&area=&numAtt=&numIns=&type=&sort=attup&view=table

Make sure you can find the required dataset before making the final decision on your project topic.

Step 3. Understanding the Data

Now you have collected data that will help you do your model. Below we will understand the variables in more details.

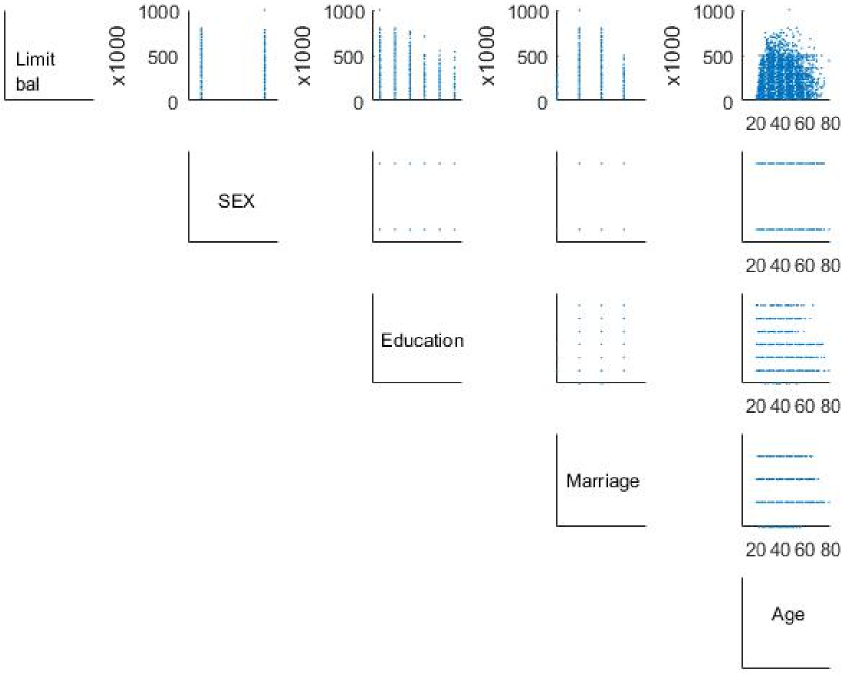
1. For each variable, find the summary statistics (mean, median, mode, minimum, maximum, variance, standard deviation, etc) using the data you collected. Create a table so that each row is one variable, and each column is one of the summary statistics.

   Make observations and write your comments.

2. Clean your data: Discuss if there are any outliers in your data or not. If there are outliers, remove them by trimming the data samples (e.g., if some houses are too cheap or too expensive comparing with the rest of houses, then removing these houses' data/information. Hint: you can use the values of standard deviation and mean to decide whether a data sample is an outlier or not).

3. Obtain scatter plots for your data. USE a scatter plot matrix to display multiple scatter plots. An example for a credit data set is given below (in this example, limit bal. is the output variable, and input variables are: sex, education, marriage, age).

   Make observations and write your comments using the scatter plots.



4. Find the correlation coefficient (the equation can be found in the statistics lecture slides) between your input variables and your output variable. Pick 3 most relevant input variables that are correlated to the output variable to use in the rest of the project.

   Explain the reasoning behind your picks. Also check if the input variables are correlated with each other.

Step 4. Feature Normalization

The idea here is to make sure input variables are on a similar scale

The simplest method is rescaling the range of each input variable to scale the range in [0, 1]. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

For example, the number of bedrooms in a house is one of the input variables you picked. x is the number of bedrooms of a particular house (say the first house in your dataset). min(x) is the smallest value of this input variable among the data set, i.e., the minimum number of bedrooms in your data set, max(x) is the largest value of this input variable among the dataset, i.e., the maximum number of bedrooms in your data set. You can find the value of x' for the first house using this equation. You have to find the x' for all the houses in the data set. Then from now on, you will use the x' values instead of x values, i.e., x' will be used as the number of bedrooms.

Step 5. Divide the data set into two groups

For each input variable and the output variable, divide the data into two groups (do the separation randomly):

a. 90 % of the data samples will be used for modeling

b. 10 % of the data samples will be used for testing

For example, if your project is the house price one, and if your dataset contains 100 houses' information (prices and house properties), 90 houses with their information will be used for modeling and 10 houses with their information will be use for testing.

Step 6. Derive the linear models to predict the output variable

We will use the linear least squares to derive models for prediction.

In brief words, we want to express your output variable as a linear combination of input variables, such as

$$\text{Output Variable } = \sum_{k=1}^{n} constant_k * input\ variable_k + constant_{n+1} \qquad \text{(Equation 1)}$$

The values of output variable and input variables are known, the objective is to find out the n+1 constant values.

Read help file for Matlab command "mldivide" and read "https://www.mathworks.com/help/matlab/data_analysis/linear-regression.html"

Make sure you use 90% of your data set (the group 1 in step 5) for modeling (finding the constants in the above equation)

Repeat this step for many models:

- Models with only one input variable (n=1; there are 3 models since you have picked three input variables in step 3.4, one model for each input variable). Find the corresponding two constant values in each model.

- Models with two input variables (n=2; there are 3 models, choose 2 input variables from three input variables, you will have 3 combinations). Find the corresponding three constant values in each model.
- Finally, models with 3 input variables (n=3; there is only 1 model). Find the corresponding three constant values in the model.

Step 7. Test the models using the test data (10% of the data samples, that is the group 2 data in step 5).

For each data sample in group 2, you will use the input variables of that data sample to calculate the output variable (predicted value) using equation 1 (recall that the constants values have been found out in Step 6). You also have the actual output variable value with that data sample (actual value). For example, if your project is house price, and if your group 2 data including 10 houses and their information, for each house, using the input variables (e.g., number of bedroom, number of bathrooms, etc) to calculate its price (using Equation 1), this is the predicted value, since you also have the house actual price, that is the actual value. So you will get 10 predicted values and 10 actual values for the 10 houses in group 2.
One of your error metrics is

$$MSE = \frac{\sum_{j=1}^{m} (actual\ value - predicted\ value)^2}{m}$$

where m is the number of data samples in your group 2.

Another metric is the R^2 value (https://www.mathworks.com/help/matlab/data_analysis/linear-regression.html).

Discuss your observations based on the error metrics. Which model performs best (recall that you have 7 models at Step 6)? Why?

Step 8. Derive a heuristic model to predict the output variable

Derive a heuristic model, f(), to predict the output variable. It could be linear or non-linear function. You should try to beat the best linear model derived in step 6. Compare the MSE for the heuristic model you derived to the best linear model derived in step 6.

$$\text{Output Variable } = f(input\ variable_1, input\ variable_2,\ input\ variable_3)$$

FINAL REPORT DUE: Dec 9th (Submit ONLINE a pdf file)

Each group submit one report.

The report should contain:

1. Title page: Project title + Group Member Names + etc.,

2. Abstract: Brief Project Description

3. List of Figures and Tables

4. Glossary: Description of Important Terms and Abbreviations used in Your Project

5. Introduction and Background: Importance of the project (the reason why you choose this project topic) + description of the other models that already exist

6. Discussion of Data

7. Analysis

8. Results

9. Conclusions (difficulties faced, any other observations/comments about the project)

10. Discussion of contribution of each team member

11. References (with proper citation in the report)

For parts 6-8, make sure you answer all the questions asked in the project. Add your comments whenever appropriate and also utilize the appropriate plots. Make sure all the figures are incorporated in the report properly.

FINAL CODE/DATA-SET DUE: Dec. 9th (Submit all your code and dataset (as an Excel file) online)

Reading Assignment:

https://en.wikipedia.org/wiki/Predictive_analytics